



IAP Sample Design, Weights, Variance Estimation, IRT Scaling and Plausible Values

Module Objectives

- Summarize the sample designs of each International Activities Program (IAP) study as they relate to study weights and describe the sampling weights that must be applied to assure data are representative of the target population
- Explain the importance of using correct techniques for variance estimation and calculating correct standard errors to be used in hypothesis testing
- Explain how scaling is used in the large-scale international assessments and how plausible values are used when analyzing assessment data

IAP Studies Sample Designs

- Not a simple random sample (SRS) of the target population
- Two-stage stratified cluster sample
 - 1st stage: selection of schools
 - Selected with probability proportional to size (PPS)
 - Assigned to strata (e.g., geographic region and school type)
 - PIRLS and TIMSS 2011
 - explicitly stratified by % of students eligible for free or reduced-price lunch, type of school, and region of the country
 - implicitly stratified by community type and minority status
 - Use of substitute schools

IAP Studies Sample Designs (Continued)

- Two-stage stratified cluster sample
 - 2nd stage ([PIRLS and TIMSS](#)): selection of classrooms within schools
 - Studies are concerned with what happens within classrooms and schools
 - 2nd stage ([PISA](#)): selection of students within schools
 - Study is concerned with 15-year-old students across grades and classes

Exclusions in IAP Study Samples

During sampling, entire schools can be excluded or specific students or entire classrooms can be excluded

- Exclusions should not exceed 5 percent of the target population
- Exclusions can only occur for specific reasons as defined (e.g., extremely small schools or students with insufficient language skills)

U.S. Exclusion Rates (Percentages)			
IAP study	School-level exclusion rate	Within-school exclusion rate	Overall exclusion rate
PIRLS 2011	0.0	7.2	7.2
TIMSS 2011			
Grade 4	0.0	7.0	7.0
Grade 8	0.0	7.2	7.2
PISA 2012	1.0	4.4	5.4

Participation Rate Standards in IAP Studies

Participation or response rate standards apply to all participating education systems

U.S. Participation Rates (Weighted percentages)				
IAP study	School participation rate before substitution	School participation rate after substitution	Student participation rate	Overall (combined) response rate with substitute schools
PIRLS 2011	80	85	96	81
TIMSS 2011				
Grade 4	79	84	95	80
Grade 8	87	87	94	81
PISA 2012	67	77	89	69

U.S. IAP Study Sample Sizes

IAP study	Number of schools	Number of students
PIRLS 2011	370	12,726
TIMSS 2011		
Grade 4	369	12,569
Grade 8	501	10,477
PISA 2012	161	6,111

General Overview of Sampling Weights

- Weights must be used to obtain correct estimates that are representative of the target population
- Weights account for the study's complex sample design, taking into account characteristics of the sample and the selection procedure
- Weights account for differential selection probabilities and nonresponse

Sampling Weights – School Weight

- Incorporates a school's probability of selection
- Accounts for sampled schools that did not participate and were not replaced
- Calculated independently for each explicit stratum

Sampling Weights – Classroom Weight (PIRLS and TIMSS)

- Reflects the probability of sampled classroom(s) being selected from among all classrooms in school at target grade level
 - Calculated independently for each school
 - Basic class-within-school weight for a sampled class is the inverse of the probability of the class being selected from all of the classes in its school
- Classroom weights are not applicable in PISA

Sampling Weights – Student Weight

- Students are assigned sampling weights to adjust for over- or under-representation of particular groups in the final sample
- Student weight is the inverse of the probability of selection
- Students with higher weight values are representing more people in the target population
- Use of sampling weights is necessary for computation of sound, nationally representative estimates
- Weights adjust for nonparticipation

Sampling Weights – Student Weight (Continued)

- Overall student sampling weights are referred to as
 - *total student weight* in PIRLS and TIMSS
 - *final student weight* in PISA
- Sum of the overall student weights equals the number of students in the target population

Sampling Weights – Teacher Weight

- Because there are no nationally representative samples of teachers (only students), analyses involving teacher data have to be viewed as student-level analyses
- Teacher weights are based on the total student weight
- A teacher questionnaire was administered in PISA starting with the 2015 administration; thus, teacher weights are not available in PISA data files prior to that time

Sampling Weights Calculated in IAP Studies

PIRLS	TIMSS	PISA
<ul style="list-style-type: none"> • School • Classroom • Student • Total Student • Teacher • House • Senate 	<ul style="list-style-type: none"> • School • Classroom • Student • Total Student • Overall Teacher • Mathematics Teacher • Science Teacher • House • Senate 	<ul style="list-style-type: none"> • School • Student • Final Student
NOTE: Teacher weights available in PISA starting with the release of PISA 2015 data.		

How to Decide Which Weight to Use

- In PISA, final student weight is most commonly used in student-level statistical analyses
- In PIRLS and TIMSS, total student weight is commonly used in student-level statistical analyses
 - Senate weight can be used for cross-country analyses in which countries should be treated equally
 - House weight ensures that weighted sample corresponds to actual sample size in each country
- Teacher weight used when using teacher data in student-level analyses
- Be cautious in use of school weights, as target population is students, not schools
- IEA IDB Analyzer automatically selects appropriate weight variable

Using Weights to Appropriately Calculate Estimates

	Rural schools	Urban schools	
	Stratum 1	Stratum 2	Total
Population	1,000	10,000	11,000
Sample	100	100	200
Sampling Weight	10	100	—

Using Weights to Appropriately Calculate Estimates (Continued)

	Rural schools	Urban schools	
	Stratum 1	Stratum 2	Total
Population	1,000	10,000	11,000
Sample	100	100	200
Sampling Weight	10	100	—
Unweighted Mean	500	600	550

$$\text{Unweighted Mean } (x) = \frac{\sum x}{n} = 550$$

Computation of unweighted mean: $((500)+(600))/(2) = 550$

Using Weights to Appropriately Calculate Estimates (Continued)

	Rural schools	Urban schools	
	Stratum 1	Stratum 2	Total
Population	1,000	10,000	11,000
Sample	100	100	200
Sampling Weight	10	100	—
Unweighted Mean	500	600	550
Weighted Mean	500	600	591

$$\text{Weighted Mean } (x) = \frac{\sum wgt \cdot x}{\sum wgt} = 591$$

Computation of weighted mean: $((500*10)+(600*100))/10+100 = 591$

Unweighted Versus Weighted Results: Example from PIRLS 2011, United States

Unweighted Percentages

NAT DERIVED RACE-COLLAPSED					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	WHITE, NOT HISPANIC	6182	48.6	49.6	49.6
	BLACK, NOT HISPANIC	1508	11.8	12.1	61.7
	HISPANIC	3330	26.2	26.7	88.5

Weighted Percentages

NAT DERIVED RACE-COLLAPSED					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	WHITE, NOT HISPANIC	1824721	50.8	51.8	51.8
	BLACK, NOT HISPANIC	420527	11.7	11.9	63.7
	HISPANIC	876632	24.4	24.9	88.6

Standard Errors and Variance Estimation

- Estimates from PIRLS, TIMSS, and PISA are not precise
- Standard errors are a measure of the precision of our estimates
- Estimation of this error is called variance estimation

Standard Errors and Variance Estimation (Continued)

- A two-stage design has more uncertainty (and generally larger standard errors) than a simple random sample of the same size
- Clustering effect - students within the same school tend to be more similar to one another on characteristics than students across all schools in the population
- In studies using a complex sample design, standard errors tend to get larger as sample sizes are smaller and when there is less variability among students within schools and more variability among students between schools

Standard Errors and Variance Estimation (Continued)

- Formulas for calculating standard errors are more complex than what is used for a simple random sample (SRS)
- Most statistical software packages assume SRS and will generate incorrect p-values in hypothesis testing
- Special statistical software is available which automatically uses sampling weights and correctly calculates standard errors. For example
 - [IEA IDB Analyzer](#)

Standard Error Calculations in PIRLS, TIMSS, and PISA: [Replication Techniques](#)

- This method calculates appropriate SEs based on differences between estimates from the full sample and a series of created subsamples (replicates)
- Select replicate weights that are associated with your main sampling weight
- PIRLS and TIMSS use a Jackknife Repeated Replication (JRR) method
 - Two variables (JKZONE) and (JKREP) contain jackknife replication information that can be used to correctly calculate standard errors
- PISA uses a Balanced Repeated Replication (BRR) method
 - There are 80 replicate weights that can be used to correctly calculate standard errors

Scaling of IAP Study Assessments: The Use of Item Response Theory (IRT)

Challenge is to develop an assessment that comprehensively covers the subject area(s) without overburdening individual students

- Many assessment items are needed
- Each student completes only a subset of items
- Using IRT scaling, student performance in an academic subject can be summarized on a common scale even when different students are administered different items

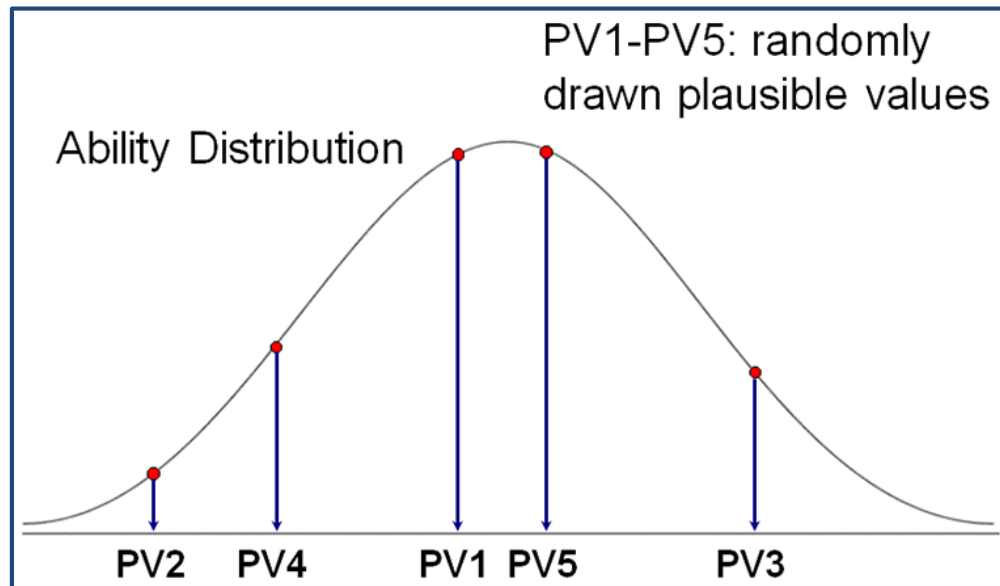
Scaling of IAP Study Assessments: The Use of IRT (Continued)

- Using IRT, we can ask, “How would the students have performed on the test had we been able to administer all of the items to all of the students?”
- IRT models allow us to create a continuum on which both student performance and item difficulty will be located, linked by a probabilistic function
- Probability of a correct answer depends on item parameters and ability of examinee
 - Students of high ability are expected to answer both easy and difficult items correctly
 - Students of low ability are not expected to answer difficult items correctly

Plausible Values Methodology

- Responses for items not completed by student must be estimated/imputed
- Plausible Values
 - Represent what true performance of student might have been, had it been observed
 - Random draws (typically 5) from the estimated ability distribution of students with similar item response patterns and background characteristics
 - Think of this as a regression where the predictors are item responses and background data
 - Variance of these draws reflects the uncertainty of measurement
 - Constructed separately for each national sample

Plausible Value Scores



Some Important Points to Bear in Mind About Plausible Values

- Always compute statistic with each plausible value and then average the results
- Both sampling variance and measurement variance must be taken into account when computing the standard errors
- Plausible values are optimal for obtaining population and subpopulation estimates
- Plausible values should not be used for individual reporting

Calculating Correct SEs for Hypothesis Testing Using [AM Statistical Software](#)

Example from TIMSS 2011, grade 8 mathematics achievement, by sex: Japan

Observations: 4414							
Weighted analysis, but assuming a simple random sample							
Sex	Weighted N	Mean	SE (Mean)	Std. Dev			
GIRL	581541	565.949	1.866	80.244			
BOY	597025	573.609	2.072	88.625			
Mean 1	Mean 2	Difference	SE Difference	Deg. of freedom	T-statistic	p > t	
565.949	573.609	-7.659	2.622	4412	-2.921	0.004	
Weighted analysis and accounting for the complex sample design of the study							
Sex	Weighted N	Mean	SE (Mean)	Std. Dev			
GIRL	581541	565.949	3.042	80.244			
BOY	597025	573.609	3.54	88.625			
Mean 1	Mean 2	Difference	SE Difference	Deg. of freedom	T-statistic	p > t	
565.949	573.609	-7.659	4.094	70	-1.871	0.066	
AM Statistical Software Beta Version 0.06.03. (c) The American Institutes for Research and Jon Cohen							

Plausible Values Example

Student ID	BSMMAT01	BSMMAT02	BSMMAT03	BSMMAT04	BSMMAT05
10454	519.32	458.32	459.32	458.32	458.32
10455	458.32	458.32	458.32	458.32	458.32
10456	522.55	459.32	500.26	495.64	465.69
10457	453.49	423.14	488.34	456.64	456.64
10458	583.91	616.97	598.23	596	
10459	526.76	469.85	456.75	481	
10460	532.03	490.87	487.52	500	
10461	532.64	511.47	499.83	515	
10462	568.74	560.02	506.35	550	
10463	489.18	511.64	533.26	461	

Variable	min	Mean	max
BSMMAT01 *1ST PLAUSIBLE VALUE MATHEMATICS*	508.9190		
BSMMAT02 *2ND PLAUSIBLE VALUE MATHEMATICS*	509.5493		
BSMMAT03 *3RD PLAUSIBLE VALUE MATHEMATICS*	510.2348		
BSMMAT04 *4TH PLAUSIBLE VALUE MATHEMATICS*	509.5759		
BSMMAT05 *5TH PLAUSIBLE VALUE MATHEMATICS*	509.5759		
Valid N (listwise)	10477		

Module Summary and Resources

- Summarized the sample designs of each International Activities Program (IAP) study as they relate to study weights and describe the sampling weights that must be applied to assure data are representative of the target population
- Explained the importance of using correct techniques for variance estimation and calculating correct standard errors to be used in hypothesis testing
- Explained how scaling is used in the large-scale international assessments and how plausible values are used when analyzing assessment data

Resources

- [Standard Errors](#)
- [Analyzing NCES Complex Survey Data](#)
- [Methods and Procedures in TIMSS and PIRLS 2011](#)
- [PISA Technical Report](#)
- [AM Statistical Software](#)
- [IEA IDB Analyzer](#)